Airline Data Analysis

Yamkela Macwili

Introduction

Delays in airline flight can be frustrating for both passengers and airlines. While some delays are caused by factors beyond anyone's control, others can be predicted and possibly prevented. To address this issue, we aim to answer the question: "can we predict the likelihood of a flight delay?".

Predicting flight delays is a critical task that can have a significant impact on the aviation industry. By identifying the factors that contribute to delays and developing predictive models, enhance passenger experience, and reduce the financial burden of delays on airlines.

In this analysis, we will explore historical flight data to uncover patterns and relationships that may help us predict flight delays. We will investigate various factors such as departure delays, carrier-related delays, and more.

Load Required Libraries

```r
library("tidyverse")
```

```
## ── Attaching core tidyverse packages ──────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────── tidyverse_conflicts()
──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force
all conflicts to become errors
```

```r
library("Hmisc")
```

```
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
##
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library("corrplot")
```

```
## corrplot 0.92 loaded
```

```
library(knitr)
```

Load and Inspect the Dataset

```
# Define the file path and read the dataset
file_path <- "C:/Users/yamke/Downloads/airline_2m/airline_2m.csv"
sub_airline <- read.csv(file_path, nrows = 8000)

# Display the structure of the dataset
glimpse(sub_airline)
```

```
## Rows: 8,000
## Columns: 109
## $ Year                            <int> 1998, 2009, 2013, 2010, 2006, 1995,
20…
## $ Quarter                         <int> 1, 2, 2, 3, 1, 4, 3, 2, 3, 1, 4, 2,
3,…
## $ Month                           <int> 1, 5, 6, 8, 1, 11, 8, 6, 8, 2, 11,
4, …
## $ DayofMonth                      <int> 2, 28, 29, 31, 15, 29, 7, 11, 3, 8,
21…
## $ DayOfWeek                       <int> 5, 4, 6, 2, 7, 3, 1, 2, 7, 4, 4, 4,
7,…
## $ FlightDate                      <chr> "1998-01-02", "2009-05-28", "2013-
06-2…
## $ Reporting_Airline               <chr> "NW", "FL", "MQ", "DL", "US", "DL",
"C…
## $ DOT_ID_Reporting_Airline        <int> 19386, 20437, 20398, 19790, 20355,
197…
## $ IATA_CODE_Reporting_Airline     <chr> "NW", "FL", "MQ", "DL", "US", "DL",
"C…
## $ Tail_Number                     <chr> "N297US", "N946AT", "N665MQ",
"N6705Y"…
## $ Flight_Number_Reporting_Airline <int> 675, 671, 3297, 1806, 465, 1198,
1431,…
## $ OriginAirportID                 <int> 13487, 13342, 11921, 12892, 11618,
112…
## $ OriginAirportSeqID              <int> 1348701, 1334202, 1192102, 1289201,
11…
## $ OriginCityMarketID              <int> 31650, 33342, 31921, 32575, 31703,
301…
## $ Origin                          <chr> "MSP", "MKE", "GJT", "LAX", "EWR",
"DF…
## $ OriginCityName                  <chr> "Minneapolis, MN", "Milwaukee, WI",
"G…
## $ OriginState                      <chr> "MN", "WI", "CO", "CA", "NJ", "TX",
"M…
## $ OriginStateFips                  <int> 27, 55, 8, 6, 34, 48, 25, 13, 17,
17, …
## $ OriginStateName                  <chr> "Minnesota", "Wisconsin",
"Colorado", …
## $ OriginWac                        <int> 63, 45, 82, 91, 21, 74, 13, 34, 41,
41…
## $ DestAirportID                    <int> 14869, 13204, 11298, 11433, 11057,
148…
## $ DestAirportSeqID                 <int> 1486902, 1320401, 1129803, 1143301,
11…
## $ DestCityMarketID                 <int> 34614, 31454, 30194, 31295, 31057,
304…
## $ Dest                             <chr> "SLC", "MCO", "DFW", "DTW", "CLT",
"SH…
## $ DestCityName                     <chr> "Salt Lake City, UT", "Orlando,
FL", "…
## $ DestState                        <chr> "UT", "FL", "TX", "MI", "NC", "LA",
"O…
## $ DestStateFips                    <int> 49, 12, 48, 26, 37, 22, 39, 45, 39,
48…
## $ DestStateName                    <chr> "Utah", "Florida", "Texas",
"Michigan"…
## $ DestWac                          <int> 87, 33, 74, 43, 36, 72, 44, 37, 44,
74…
## $ CRSDepTime                       <int> 1640, 1204, 1630, 1305, 1820, 639,
175…
## $ DepTime                          <int> 1659, 1202, 1644, 1305, 1911, 639,
175…
## $ DepDelay                         <dbl> 19, -2, 14, 0, 51, 0, -4, 221, 2,
16, …
## $ DepDelayMinutes                  <dbl> 19, 0, 14, 0, 51, 0, 0, 221, 2, 16,
2,…
## $ DepDel15                         <dbl> 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0,
0,…
## $ DepartureDelayGroups             <int> 1, -1, 0, 0, 3, 0, -1, 12, 0, 1, 0,
0,…
## $ DepTimeBlk                       <chr> "1600-1659", "1200-1259", "1600-
1659",…
## $ TaxiOut                          <dbl> 24, 10, 9, 23, 19, 29, 33, 19, 26,
34,…
## $ WheelsOff                        <int> 1723, 1212, 1653, 1328, 1930, 708,
182…
## $ WheelsOn                         <int> 1856, 1533, 1936, 2008, 2050, 736,
195…
## $ TaxiIn                           <dbl> 3, 8, 6, 7, 8, 5, 4, 6, 3, 5, NA,
3, N…
## $ CRSArrTime                       <int> 1836, 1541, 1945, 2035, 2026, 730,
200…
## $ ArrTime                         <int> 1859, 1541, 1942, 2015, 2058, 741,
200…
## $ ArrDelay                        <dbl> 23, 0, -3, -20, 32, 11, 2, 214, 10,
29…
## $ ArrDelayMinutes                 <dbl> 23, 0, 0, 0, 32, 11, 2, 214, 10,
29, 6…
## $ ArrDel15                        <dbl> 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0,
0,…
## $ ArrivalDelayGroups              <int> 1, 0, -1, -2, 2, 0, 0, 12, 0, 1, 0,
-1…
## $ ArrTimeBlk                      <chr> "1800-1859", "1500-1559", "1900-
1959",…
## $ Cancelled                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,…
## $ CancellationCode                <chr> "", "", "", "", "", "", "", "", "",
""…
## $ Diverted                        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,…
## $ CRSElapsedTime                  <dbl> 176, 157, 135, 270, 126, 51, 125,
67, …
## $ ActualElapsedTime               <dbl> 180, 159, 118, 250, 107, 62, 131,
60, …
## $ AirTime                         <dbl> 153, 141, 103, 220, 80, 28, 94, 35,
59…
## $ Flights                         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,…
## $ Distance                        <dbl> 991, 1066, 773, 1979, 529, 190,
563, 1…
## $ DistanceGroup                   <int> 4, 5, 4, 8, 3, 1, 3, 1, 2, 4, 1, 3,
5,…
## $ CarrierDelay                    <dbl> NA, NA, NA, NA, 0, NA, NA, 0, NA,
0, N…
## $ WeatherDelay                    <dbl> NA, NA, NA, NA, 0, NA, NA, 0, NA,
0, N…
## $ NASDelay                        <dbl> NA, NA, NA, NA, 0, NA, NA, 0, NA,
13, …
## $ SecurityDelay                   <dbl> NA, NA, NA, NA, 0, NA, NA, 0, NA,
0, N…
## $ LateAircraftDelay               <dbl> NA, NA, NA, NA, 32, NA, NA, 214,
NA, 1…
## $ FirstDepTime                    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ TotalAddGTime                   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ LongestAddGTime                 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ DivAirportLandings              <int> NA, 0, 0, 0, NA, NA, NA, 0, NA, 0,
NA,…
## $ DivReachedDest                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ DivActualElapsedTime            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ DivArrDelay                     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ DivDistance                     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div1Airport                     <chr> "", "", "", "", "", "", "", "", "",
""…
## $ Div1AirportID                   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div1AirportSeqID                <int> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div1WheelsOn                    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div1TotalGTime                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div1LongestGTime                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div1WheelsOff                   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div1TailNum                     <chr> "", "", "", "", "", "", "", "", "",
""…
## $ Div2Airport                     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div2AirportID                   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div2AirportSeqID                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div2WheelsOn                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div2TotalGTime                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div2LongestGTime                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div2WheelsOff                   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div2TailNum                     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div3Airport                     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div3AirportID                   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div3AirportSeqID                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div3WheelsOn                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div3TotalGTime                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div3LongestGTime                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div3WheelsOff                   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div3TailNum                     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div4Airport                     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div4AirportID                   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div4AirportSeqID                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div4WheelsOn                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div4TotalGTime                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div4LongestGTime                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div4WheelsOff                   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div4TailNum                     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div5Airport                     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div5AirportID                   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div5AirportSeqID                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div5WheelsOn                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div5TotalGTime                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div5LongestGTime                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div5WheelsOff                   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
## $ Div5TailNum                     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA…
```

The dataset has 8000 rows and 109 columns.

Dataset Glossary

Now let’s look at the features and description of the dataset.

```
url <- "C:/Users/yamke/OneDrive/Documents/Airline
Analysis/Airline/table3.csv"
dataset_glossary <- read.csv(url)

# Print the dataset glossary in tabular format
kable(dataset_glossary)
```

Feature	Description
Year	Year
Quarter	Quarter
Month	Month
DayofMonth	Day of Month
DayOfWeek	Day of Week (numeric)
FlightDate	Date of Flight
Reporting_Airline	Airline Unique Carrier Code
DOT_ID_Reporting_Airline	Number assigned by US DOT to identify a unique airline
IATA_CODE_Reporting_Airline	Airline Code assigned by IATA
Tail_Number	Aircraft tail number
Flight_Number_Reporting_Airline	Flight Number
OriginAirportID	Origin Airport ID
OriginAirportSeqID	Origin Airport Sequence ID
OriginCityMarketID	Origin City Market ID
Origin	Origin Airport Code
OriginCityName	Origin City Name
OriginState	Origin State
OriginStateFips	Origin State FIPS place code
OriginStateName	Origin State Name
OriginWac	Origin Airport World Area Code
DestAirportID	Destination Airport ID
DestAirportSeqID	Destination Airport Sequence ID
DestCityMarketID	Destination City Market ID
Dest	Destination Airport Code
DestCityName	Destination City Name
DestState	Destination State
DestStateFips	Destination State FIPS code
DestStateName	Destination State Name
DestWac	Destination Airport World Area Code
CRSDepTime	Computer Reservation System (scheduled) Departure Time
DepTime	Departure Time (hhmm)
DepDelay	Departure delay (minutes)
DepDelayMinutes	Absolute value of DepDelay
DepDel15	Departure Delay >15?

DepartureDelayGroups	Departure delay 15 minute interval group
DepTimeBlk	Computer Reservation System (scheduled) time block
TaxiOut	Taxi out time (minutes)
WheelsOff	Wheels off time (local time, hhmm)
WheelsOn	Wheels on time (local time hhmm)
TaxiIn	Taxi in time (minutes)
CRSArrTime	Computer Reservation System (scheduled) Arrival Time
ArrTime	Arrival time (local time, hhmm)
ArrDelay	Arrival delay (minutes)
ArrDelayMinutes	Absolute value of ArrDelay
ArrDel15	Arrival Delay >15?
ArrivalDelayGroups	Arrival delay 15 minute interval group
ArrTimeBlk	Computer Reservation System (scheduled) arrival time block
Cancelled	1 = canceled
CancellationCode	A = Carrier, B = Weather, C = National Air System, D = Security
Diverted	1 = diverted
CRSElapsedTime	Computer Reservation System (scheduled) elapsed time
ActualElapsedTime	Actual elapsed time
AirTime	Flight time (minutes)
Flights	Number of flights
Distance	Distance between airports (miles)
DistanceGroup	250 mile distance interval group
CarrierDelay	Carrier delay (minutes)
WeatherDelay	Weather delay (minutes)
NASDelay	National Air System delay (minutes)
SecurityDelay	Security delay (minutes)
LateAircraftDelay	Late aircraft delay (minutes)
FirstDepTime	First gate departure time at origin airport
TotalAddGTime	Total ground time away from gate
LongestAddGTime	Longest time away from gate
DivAirportLandings	Number of diverted airport landings
DivReachedDest	1 = diverted flight reached scheduled destination

DivActualElapsedTime	Elapsed time of diverted flight reaching scheduled destination
DivArrDelay	Difference in minutes between scheduled and actual arrival time
DivDistance	Distance between scheduled and diverted airport
Div1Airport	Diverted Airport 1
Div1AirportID	Diverted Airport 1 ID
Div1AirportSeqID	Diverted Airport 1 Sequence ID
Div1WheelsOn	Diverted Airport 1 wheels on time (local, hhmm)
Div1TotalGTime	Diverted Airport 1 total ground time away from gate
Div1LongestGTime	Diverted Airport 1 longest ground time away from gate
Div1WheelsOff	Diverted Airport 1 wheels off time (local, hhmm)
Div1TailNum	Diverted Airport 1 aircraft tail number
Div2Airport	Diverted Airport 2
Div2AirportID	Diverted Airport 2 ID
Div2AirportSeqID	Diverted Airport 2 Sequence ID
Div2WheelsOn	Diverted Airport 2 wheels on time (local, hhmm)
Div2TotalGTime	Diverted Airport 2 total ground time away from gate
Div2LongestGTime	Diverted Airport 2 longest ground time away from gate
Div2WheelsOff	Diverted Airport 2 wheels off time (local, hhmm)
Div2TailNum	Diverted Airport 2 aircraft tail number
Div3Airport	Diverted Airport 3
Div3AirportID	Diverted Airport 3 ID
Div3AirportSeqID	Diverted Airport 3 Sequence ID
Div3WheelsOn	Diverted Airport 3 wheels on time (local, hhmm)
Div3TotalGTime	Diverted Airport 3 total ground time away from gate
Div3LongestGTime	Diverted Airport 3 longest ground time away from gate
Div3WheelsOff	Diverted Airport 3 wheels off time (local, hhmm)
Div3TailNum	Diverted Airport 3 aircraft tail number
Div4Airport	Diverted Airport 4
Div4AirportID	Diverted Airport 4 ID
Div4AirportSeqID	Diverted Airport 4 Sequence ID

Div4WheelsOn	Diverted Airport 4 wheels on time (local, hhmm)
Div4TotalGTime	Diverted Airport 4 total ground time away from gate
Div4LongestGTime	Diverted Airport 4 longest ground time away from gate
Div4WheelsOff	Diverted Airport 4 wheels off time (local, hhmm)
Div4TailNum	Diverted Airport 4 aircraft tail number
Div5Airport	Diverted Airport 5
Div5AirportID	Diverted Airport 5 ID
Div5AirportSeqID	Diverted Airport 5 Sequence ID
Div5WheelsOn	Diverted Airport 5 wheels on time (local, hhmm)
Div5TotalGTime	Diverted Airport 5 total ground time away from gate
Div5LongestGTime	Diverted Airport 5 longest ground time away from gate
Div5WheelsOff	Diverted Airport 5 wheels off time (local, hhmm)
Div5TailNum	Diverted Airport 5 aircraft tail number

Variables of the Airline-Reporting Carrier On-Time Performance dataset include:

- The target value “ArrDelay” or “ArrDelayMinutes”.
- Reasons for delay.

Data Preprocessing

Lets us have a look at the our dataset for missing values.

```r
# Count missing values in all columns
missing_values <- sub_airline %>% map(~sum(is.na(.)))
missing_values
```

```
## $Year
## [1] 0
##
## $Quarter
## [1] 0
##
## $Month
## [1] 0
##
## $DayofMonth
## [1] 0
##
## $DayOfWeek
## [1] 0
```

```
##
## $FlightDate
## [1] 0
##
## $Reporting_Airline
## [1] 0
##
## $DOT_ID_Reporting_Airline
## [1] 0
##
## $IATA_CODE_Reporting_Airline
## [1] 0
##
## $Tail_Number
## [1] 0
##
## $Flight_Number_Reporting_Airline
## [1] 0
##
## $OriginAirportID
## [1] 0
##
## $OriginAirportSeqID
## [1] 0
##
## $OriginCityMarketID
## [1] 0
##
## $Origin
## [1] 0
##
## $OriginCityName
## [1] 0
##
## $OriginState
## [1] 0
##
## $OriginStateFips
## [1] 1
##
## $OriginStateName
## [1] 0
##
## $OriginWac
## [1] 0
##
## $DestAirportID
## [1] 0
##
## $DestAirportSeqID
## [1] 0
##
## $DestCityMarketID
## [1] 0
##
## $Dest
## [1] 0
##
## $DestCityName
## [1] 0
##
## $DestState
## [1] 0
##
## $DestStateFips
## [1] 2
##
## $DestStateName
## [1] 0
##
## $DestWac
## [1] 0
##
## $CRSDepTime
## [1] 0
##
## $DepTime
## [1] 133
##
## $DepDelay
## [1] 133
##
## $DepDelayMinutes
## [1] 133
##
## $DepDel15
## [1] 133
##
## $DepartureDelayGroups
## [1] 133
##
## $DepTimeBlk
## [1] 0
##
## $TaxiOut
## [1] 1661
##
## $WheelsOff
## [1] 1661
##
```

```
## $WheelsOn
## [1] 1672
##
## $TaxiIn
## [1] 1672
##
## $CRSArrTime
## [1] 0
##
## $ArrTime
## [1] 147
##
## $ArrDelay
## [1] 152
##
## $ArrDelayMinutes
## [1] 152
##
## $ArrDel15
## [1] 152
##
## $ArrivalDelayGroups
## [1] 152
##
## $ArrTimeBlk
## [1] 0
##
## $Cancelled
## [1] 0
##
## $CancellationCode
## [1] 0
##
## $Diverted
## [1] 0
##
## $CRSElapsedTime
## [1] 2
##
## $ActualElapsedTime
## [1] 152
##
## $AirTime
## [1] 1677
##
## $Flights
## [1] 0
##
## $Distance
## [1] 0
```

```
##
## $DistanceGroup
## [1] 0
##
## $CarrierDelay
## [1] 7136
##
## $WeatherDelay
## [1] 7136
##
## $NASDelay
## [1] 7136
##
## $SecurityDelay
## [1] 7136
##
## $LateAircraftDelay
## [1] 7136
##
## $FirstDepTime
## [1] 7990
##
## $TotalAddGTime
## [1] 7990
##
## $LongestAddGTime
## [1] 7990
##
## $DivAirportLandings
## [1] 5055
##
## $DivReachedDest
## [1] 7993
##
## $DivActualElapsedTime
## [1] 7995
##
## $DivArrDelay
## [1] 7995
##
## $DivDistance
## [1] 7993
##
## $Div1Airport
## [1] 0
##
## $Div1AirportID
## [1] 7992
##
## $Div1AirportSeqID
## [1] 7992
##
## $Div1WheelsOn
## [1] 7992
##
## $Div1TotalGTime
## [1] 7992
##
## $Div1LongestGTime
## [1] 7992
##
## $Div1WheelsOff
## [1] 7995
##
## $Div1TailNum
## [1] 0
##
## $Div2Airport
## [1] 8000
##
## $Div2AirportID
## [1] 8000
##
## $Div2AirportSeqID
## [1] 8000
##
## $Div2WheelsOn
## [1] 8000
##
## $Div2TotalGTime
## [1] 8000
##
## $Div2LongestGTime
## [1] 8000
##
## $Div2WheelsOff
## [1] 8000
##
## $Div2TailNum
## [1] 8000
##
## $Div3Airport
## [1] 8000
##
## $Div3AirportID
## [1] 8000
##
## $Div3AirportSeqID
## [1] 8000
##
```

```
## $Div3WheelsOn
## [1] 8000
##
## $Div3TotalGTime
## [1] 8000
##
## $Div3LongestGTime
## [1] 8000
##
## $Div3WheelsOff
## [1] 8000
##
## $Div3TailNum
## [1] 8000
##
## $Div4Airport
## [1] 8000
##
## $Div4AirportID
## [1] 8000
##
## $Div4AirportSeqID
## [1] 8000
##
## $Div4WheelsOn
## [1] 8000
##
## $Div4TotalGTime
## [1] 8000
##
## $Div4LongestGTime
## [1] 8000
##
## $Div4WheelsOff
## [1] 8000
##
## $Div4TailNum
## [1] 8000
##
## $Div5Airport
## [1] 8000
##
## $Div5AirportID
## [1] 8000
##
## $Div5AirportSeqID
## [1] 8000
##
## $Div5WheelsOn
## [1] 8000
```

```
##
## $Div5TotalGTime
## [1] 8000
##
## $Div5LongestGTime
## [1] 8000
##
## $Div5WheelsOff
## [1] 8000
##
## $Div5TailNum
## [1] 8000
```

Dealing with the missing values first.

N/A on our delay type means no delay, therefore we replace the missing values with 0.

```
#Replace the missing values
sub_airline <- sub_airline %>%
  mutate(across(contains("Delay"), ~replace(., is.na(.), 0)))

# Select and view the delay-related columns
delay_columns <- select(sub_airline, contains("Delay"))
kable(head(delay_columns))
```

DepDelay	DepDelayMinutes	DepartureDelayGroups	ArrDelay	ArrDelayMinutes	ArrivalDelayGroups	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay	LateAircraftDelay	DivArrDelay
19	19	1	23	23	1	0	0	0	0	0	0
-2	0	-1	0	0	0	0	0	0	0	0	0
14	14	0	-3	0	-1	0	0	0	0	0	0
0	0	0	-20	0	-2	0	0	0	0	0	0
51	51	3	32	32	2	0	0	0	0	32	0
0	0	0	11	11	0	0	0	0	0	0	0

All the missing values are replaced with 0. Now we can visualize the data.

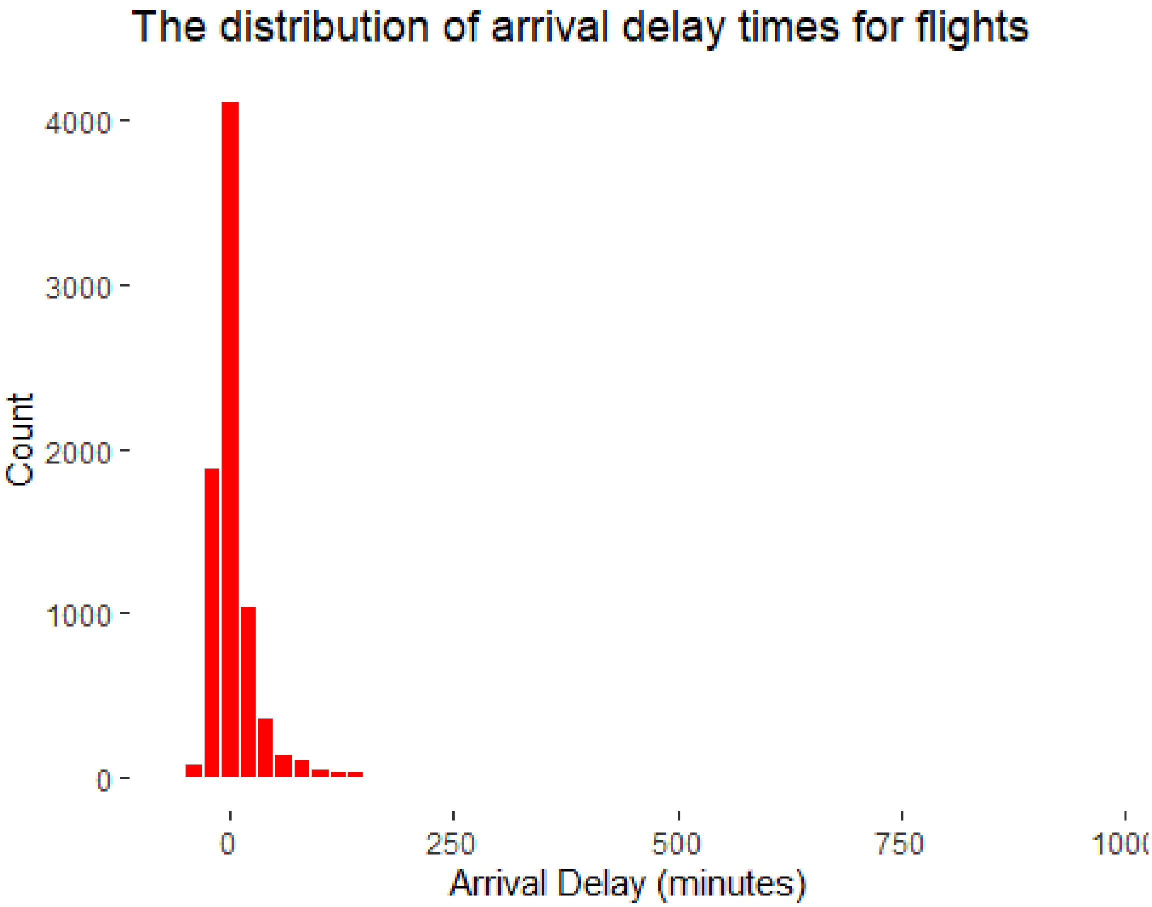
Data Visualization

Histogram of ArrDelay

```
# Check the range
range(sub_airline$ArrDelay)
```

```
## [1] -60 954
```

```
# Create a histogram of ArrDelay
ggplot(data = sub_airline, mapping = aes(x = ArrDelay)) +
  geom_histogram(binwidth = 20, color = "white", fill = "red") +
  coord_cartesian(xlim = c(-60, 954))+
  labs(title = "The distribution of arrival delay times for flights",
       x = "Arrival Delay (minutes)", y = "Count")
```



What causes a flight delay?

Let’s see how many flights are associated with each unique reporting airline.

```
count_airline <- sub_airline%>%
  count(sub_airline$Reporting_Airline)
kable(count_airline)
```

sub_airline$Reporting_Airline	n
9E	83
AA	942
AS	178
B6	118
CO	330
DH	26

DL	1079
EA	30
EV	274
F9	60
FL	99
G4	5
HA	38
HP	157
KH	4
ML (1)	3
MQ	330
NK	40
NW	432
OH	90
OO	461
PA (1)	14
PI	45
PS	3
TW	147
TZ	15
UA	758
US	723
VX	16
WN	1239
XE	131
YV	99
YX	31

```r
# Descriptive Statistics - Mean and Standard Deviation of ArrDelayMinutes
summary_airline_delay <- sub_airline %>%
  group_by(Reporting_Airline) %>%
  summarise(mean = mean(ArrDelayMinutes), std_dev = sd(ArrDelayMinutes))
kable(summary_airline_delay)
```

Reporting_Airline	mean	std_dev
9E	15.578313	37.349085
AA	11.145435	27.128910
AS	11.247191	28.045476
B6	11.262712	27.922796

CO	12.866667	31.748161
DH	10.461538	21.628187
DL	10.142725	27.598164
EA	10.933333	18.754187
EV	13.645985	49.372423
F9	28.733333	85.276195
FL	11.050505	24.889705
G4	6.800000	7.563068
HA	3.947368	7.986297
HP	13.057325	44.700032
KH	12.250000	21.203380
ML (1)	3.666667	6.350853
MQ	11.215151	25.902185
NK	14.925000	29.452602
NW	9.104167	21.059650
OH	18.166667	43.053207
OO	15.477223	60.050758
PA (1)	5.928571	11.605465
PI	11.200000	15.862477
PS	7.333333	12.701706
TW	12.231293	30.210691
TZ	7.733333	25.038447
UA	16.875989	42.584539
US	10.201936	23.761118
VX	23.562500	74.635978
WN	10.263922	24.938320
XE	11.847328	32.571454
YV	13.373737	39.244806
YX	23.774193	100.237621

```r
# Create a simple average across Reporting_Airline and DayOfWeek
average_delays <- sub_airline %>%
  group_by(Reporting_Airline, DayOfWeek)%>%
  summarise(mean_delays = mean(ArrDelayMinutes))
```

```
## `summarise()` has grouped output by 'Reporting_Airline'. You can override
using
## the `.groups` argument.
```

```r
kable(head(average_delays))
```

Reporting_Airline	DayOfWeek	mean_delays
9E	1	8.473684
9E	2	29.545455
9E	3	18.619048
9E	4	4.333333
9E	5	18.933333
9E	6	3.857143

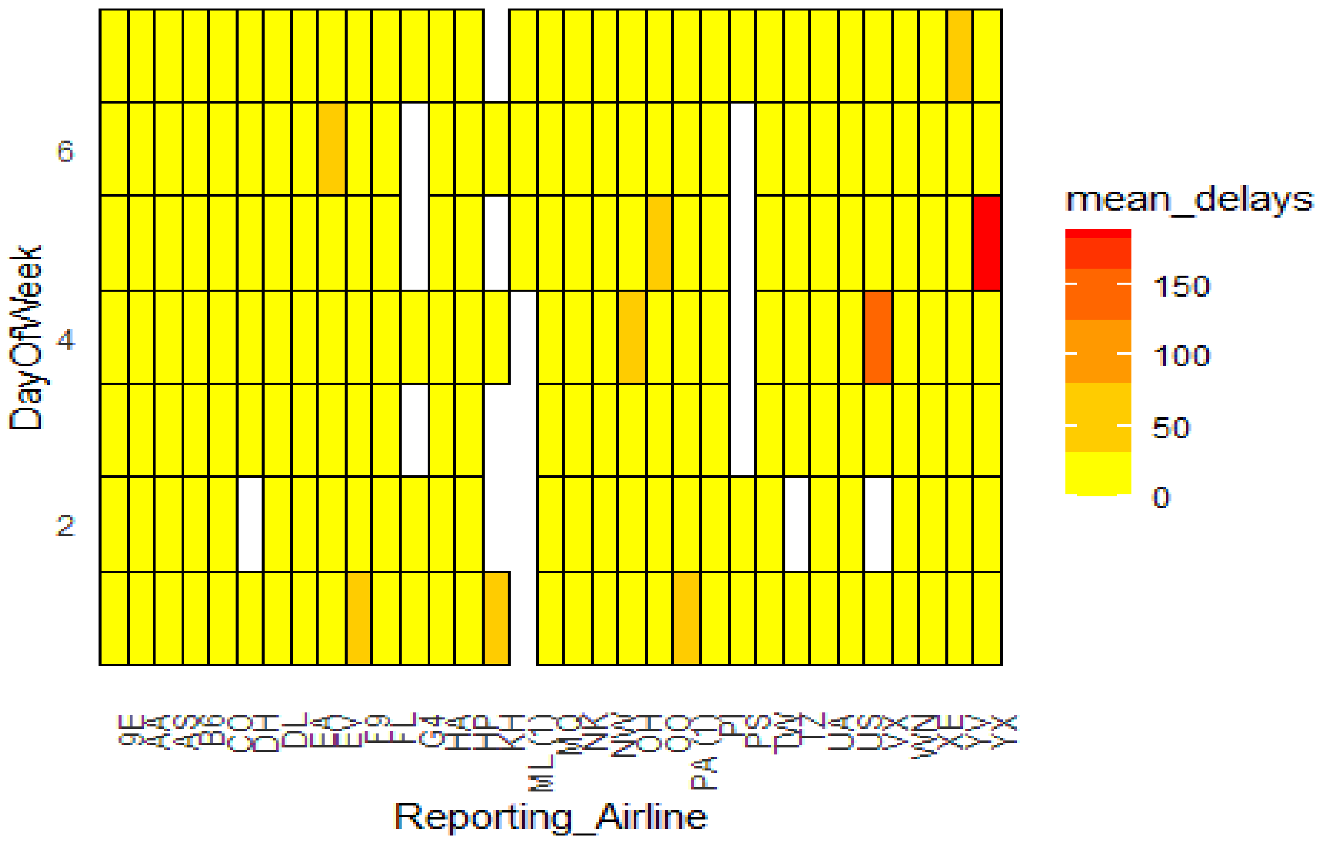
```r
#Sort the dataframe
arrange_avg_delay <- average_delays %>%
  arrange(desc(mean_delays))
kable(head(arrange_avg_delay))
```

Reporting_Airline	DayOfWeek	mean_delays
YX	5	187.33333
VX	4	150.50000
F9	1	74.63636
OH	4	46.63636
KH	1	44.00000
OO	5	39.97015

Visualize the data

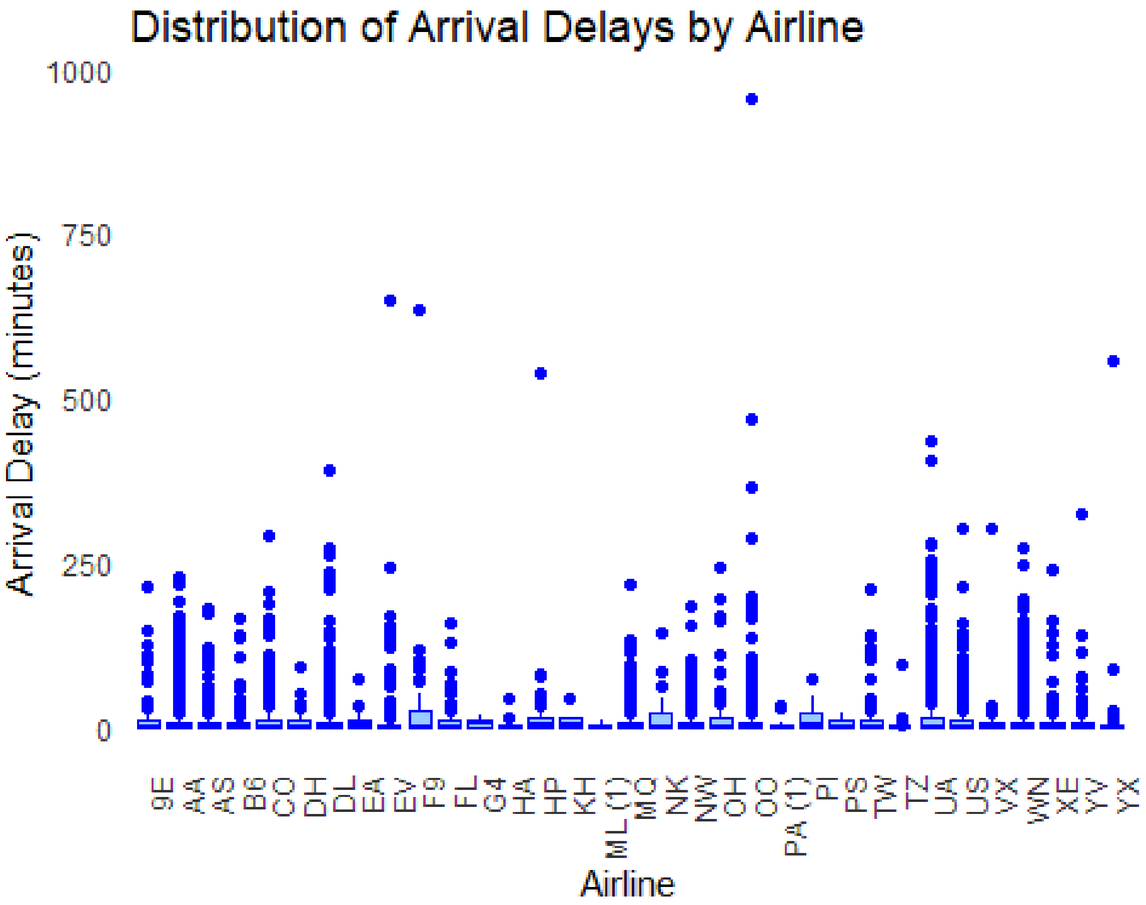
Heatmap

```r
# Visualize the data using Heatmap
average_delays %>%
  ggplot(aes( x = Reporting_Airline,
              y = DayOfWeek,
              fill = mean_delays))+
  geom_tile(color = "black", linewidth = 0.2)+
  scale_fill_gradient(low = "yellow",
                      high = "red")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90 , hjust = 1))
```



Box plot

```r
ggplot(data = sub_airline, mapping = aes(x = Reporting_Airline, y =
ArrDelayMinutes))+
  geom_boxplot(fill = "lightblue", color = "blue")+
  labs(title = "Distribution of Arrival Delays by Airline",
       x = "Airline", y = "Arrival Delay (minutes)")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, hjust =1))
```



Linear Relationships

Correlation Matrix

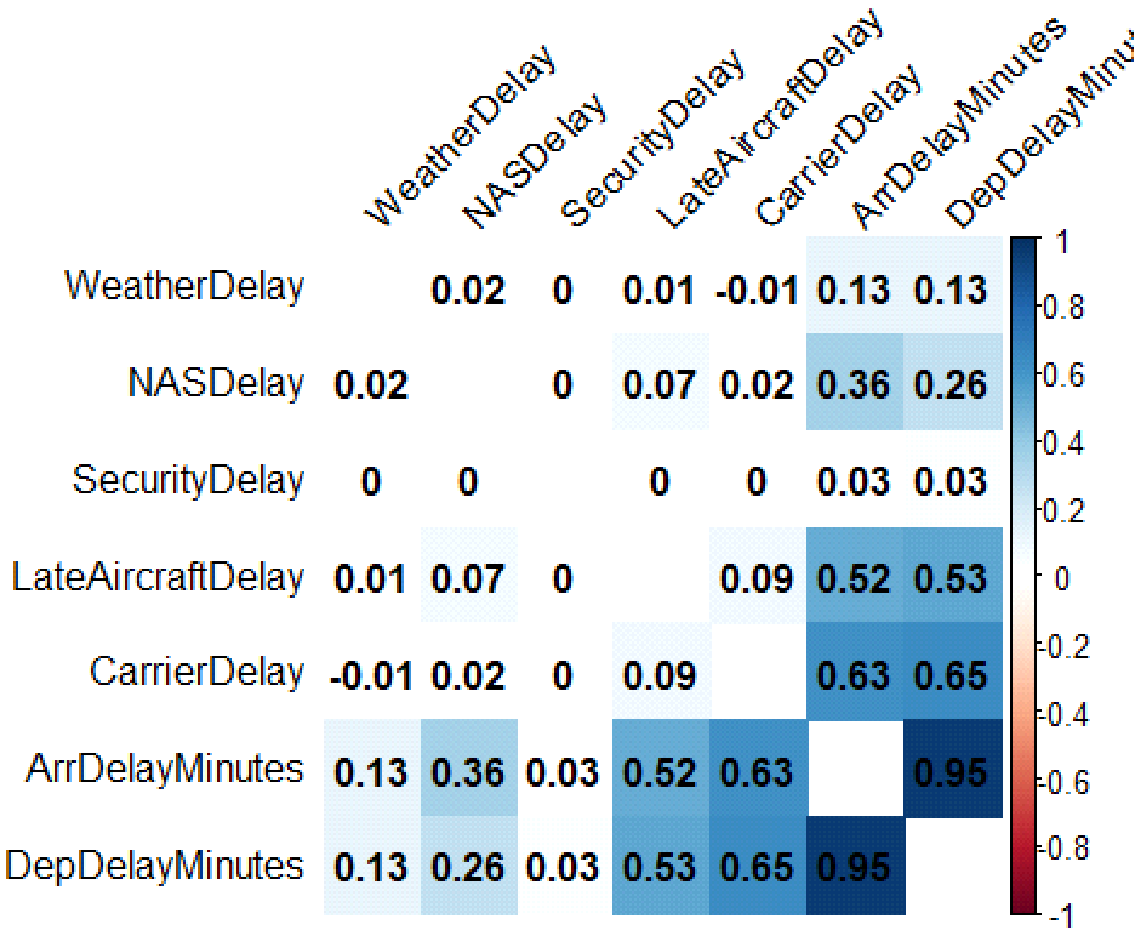
```
# Correlation between different delay types and ArrDelayMinutes
corr_airline <- sub_airline %>%
  select(ArrDelayMinutes, DepDelayMinutes, CarrierDelay, WeatherDelay,
NASDelay, SecurityDelay, LateAircraftDelay)

airline_correlation <- rcorr(as.matrix(corr_airline), type = "pearson")
correlation_matrix <- airline_correlation$r
kable(correlation_matrix, format = "markdown")
```

	ArrDelayMinutes	DepDelayMinutes	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay	LateAircraftDelay
ArrDelayMinutes	1.0000000	0.9510640	0.6266076	0.1293258	0.3583874	0.0289321	0.5169887
DepDelayMinutes	0.9510640	1.0000000	0.6484311	0.1291708	0.2634974	0.0315422	0.5303645
CarrierDelay	0.6266076	0.6484311	1.0000000	-0.00533	0.0180580	-0.00203	0.0888572

				10		72	
WeatherDelay	0.1293258	0.1291708	-0.0053310	1.0000000	0.0224852	-0.0010936	0.0149123
NASDelay	0.3583874	0.2634974	0.0180580	0.0224852	1.0000000	-0.0028939	0.0731567
SecurityDelay	0.0289321	0.0315422	-0.0020372	-0.0010936	-0.0028939	1.0000000	0.0013452
LateAircraftDelay	0.5169887	0.5303645	0.0888572	0.0149123	0.0731567	0.0013452	1.0000000

Correlation Heatmap

```
corrplot(correlation_matrix, method = "color", tl.col = "black", tl.srt = 45,
addCoef.col = "black", type = "full",
         diag = FALSE, order = "hclust")
```



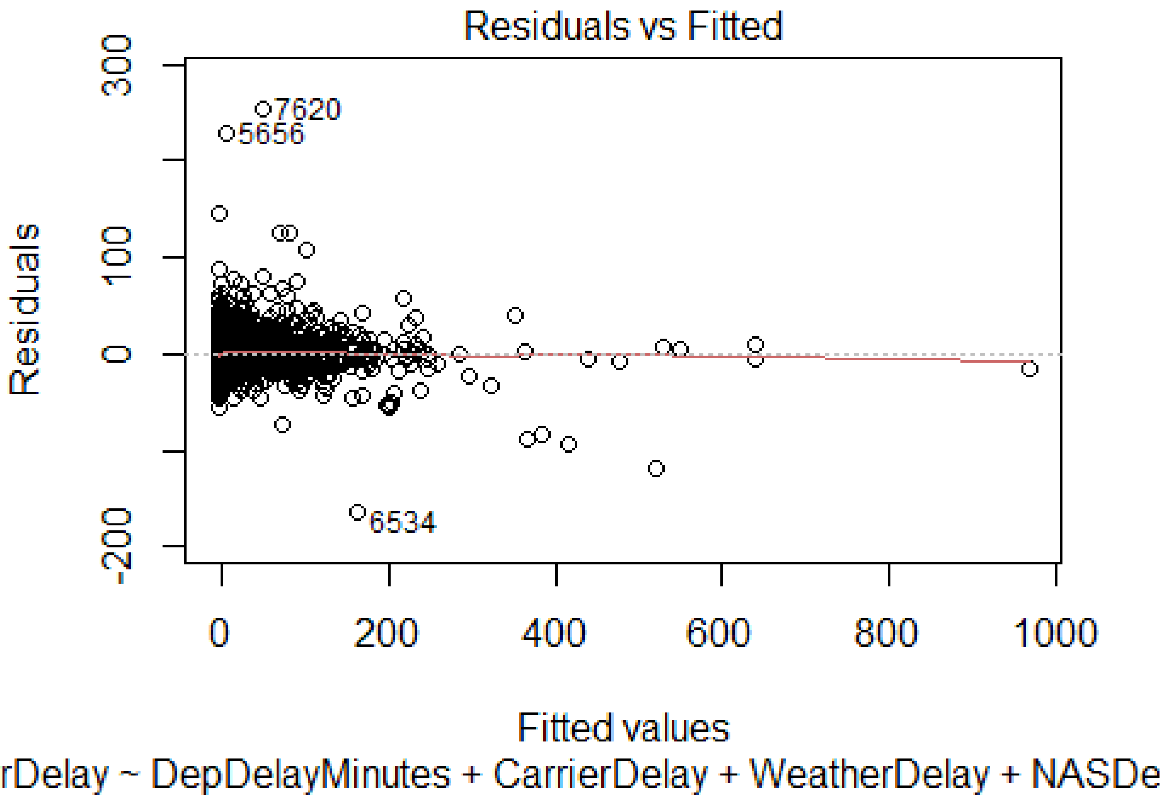
Model development

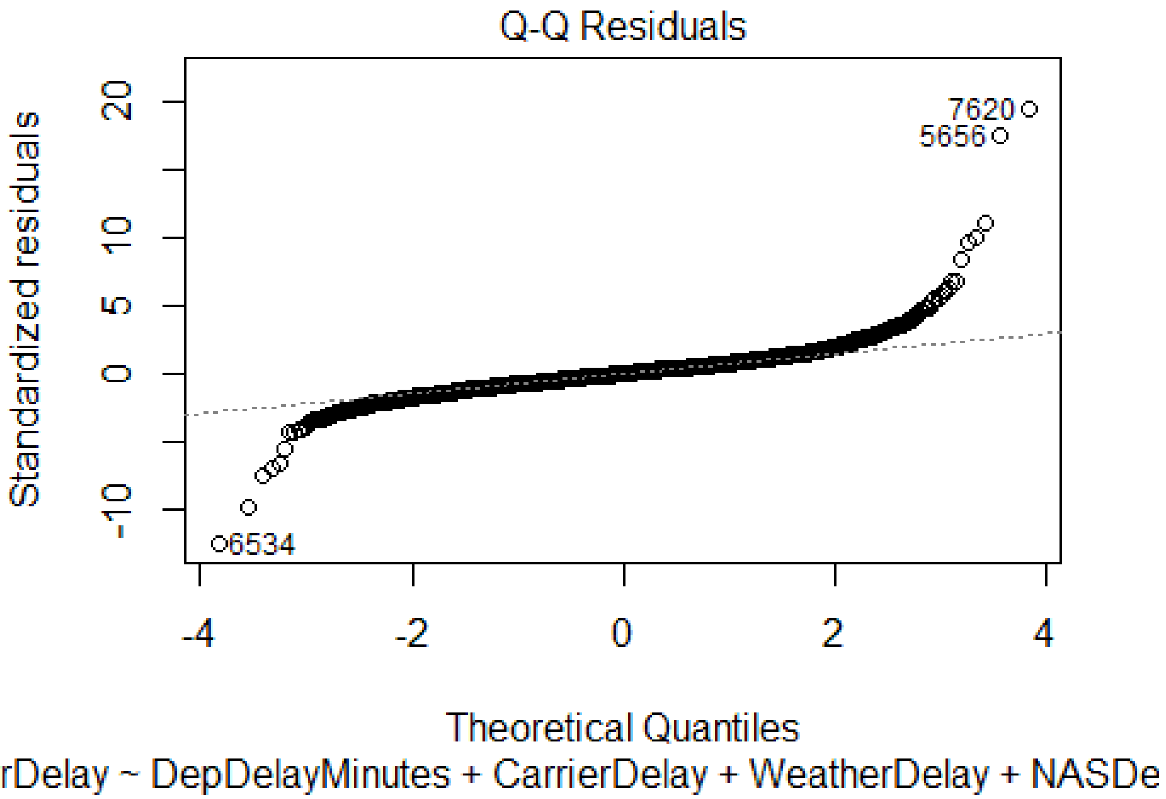
Can we predict the arrival delay of a flight?

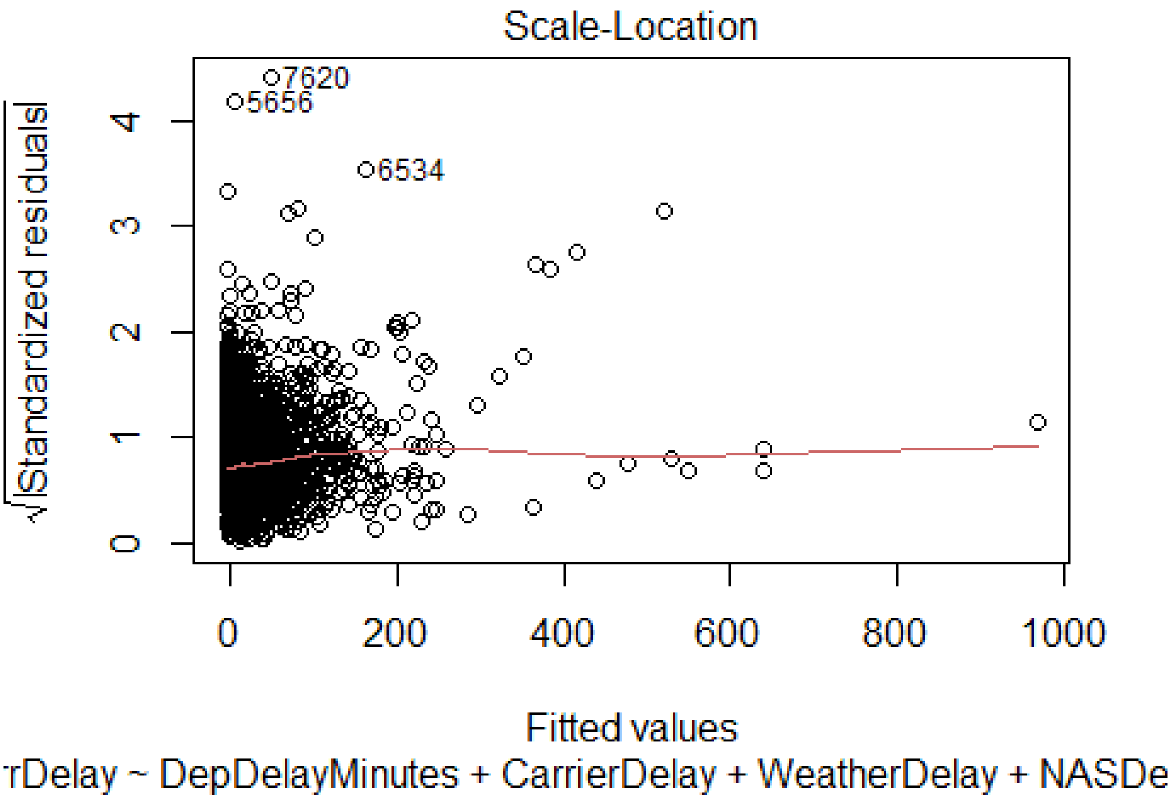
```
mult_linear_reg <- lm(ArrDelay ~ DepDelayMinutes + CarrierDelay +
WeatherDelay + NASDelay + SecurityDelay + LateAircraftDelay, data =
sub_airline)
summary(mult_linear_reg)
```

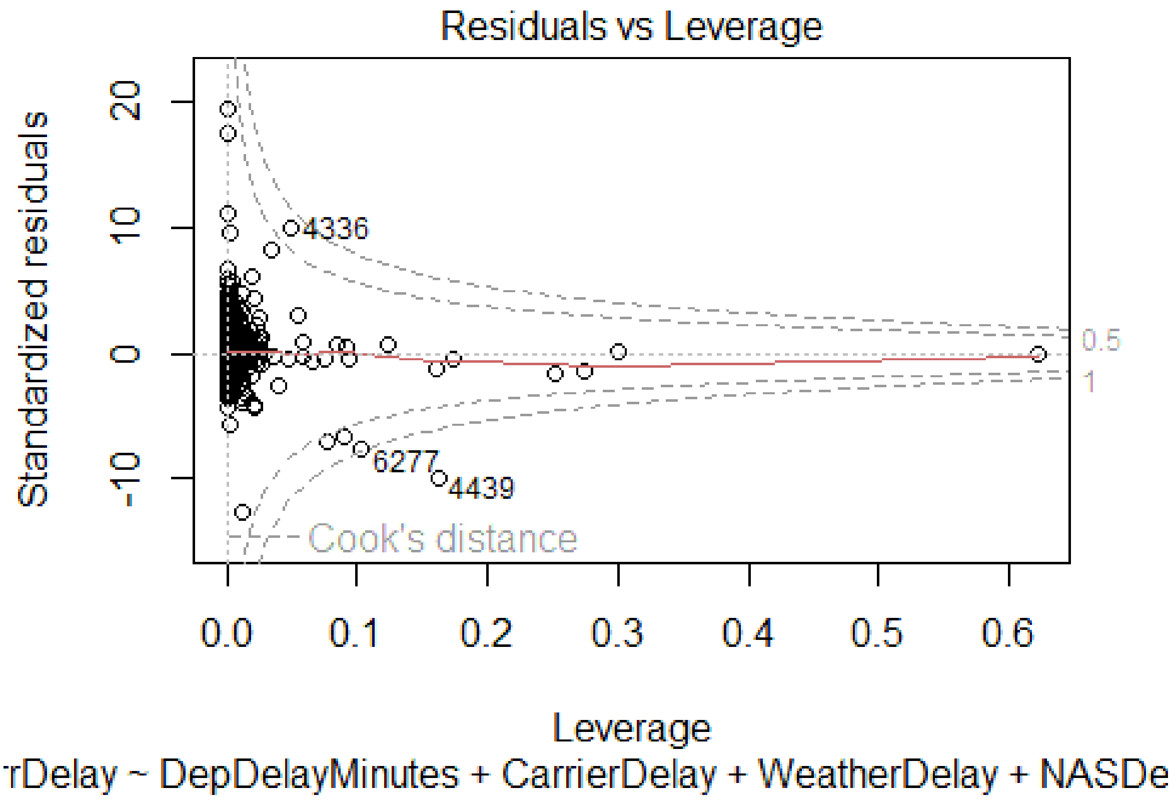
```
## 
## Call:
## lm(formula = ArrDelay ~ DepDelayMinutes + CarrierDelay + WeatherDelay + 
##     NASDelay + SecurityDelay + LateAircraftDelay, data = sub_airline)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -162.674   -6.923   -0.391    6.313  253.618 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -4.522694   0.154887 -29.200  < 2e-16 ***
## DepDelayMinutes   0.913642   0.007986 114.402  < 2e-16 ***
## CarrierDelay      0.086193   0.010719   8.041 1.02e-15 ***
## WeatherDelay      0.140914   0.034342   4.103 4.11e-05 ***
## NASDelay          0.414829   0.013742  30.187  < 2e-16 ***
## SecurityDelay     0.123307   0.125513   0.982    0.326    
## LateAircraftDelay 0.106597   0.012083   8.822  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 13.03 on 7993 degrees of freedom
## Multiple R-squared:  0.8756, Adjusted R-squared:  0.8755 
## F-statistic:  9379 on 6 and 7993 DF,  p-value: < 2.2e-16
```

```
plot(mult_linear_reg)
```


Residuals vs Fitted
Residuals
300
100
0
-200
7620
5656
6534
0
200
400
600
800
1000
Fitted values
rDelay ~ DepDelayMinutes + CarrierDelay + WeatherDelay + NASDe


Q-Q Residuals
Standardized residuals
20
10
5
0
-10
7620
5656
6534
-4
-2
0
2
4
Theoretical Quantiles
rDelay ~ DepDelayMinutes + CarrierDelay + WeatherDelay + NASDe


Scale-Location
7620
5656
6534
√|Standardized residuals|
0
1
2
3
4
0
200
400
600
800
1000
Fitted values
rDelay ~ DepDelayMinutes + CarrierDelay + WeatherDelay + NASDe


Residuals vs Leverage
Standardized residuals
20
10
0
-10
4336
6277
4439
Cook's distance
0.5
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Leverage
rDelay ~ DepDelayMinutes + CarrierDelay + WeatherDelay + NASDe